Variational Autoencoders for Biosensor Data Augmentation

Solomon Kim and Rodney Summerscales Ph.D.

Department of Computing - College of Professions, Andrews University

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.

Methodology

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 on every concentration. In each epoch we randomly sampled 10 times from the selected current samples. This allowed us to have a varied breadth of training samples for each concentration. Each time we computed the loss and optimized the model accordingly. We then tested on a subset of these values in order to determine our ELBO for that epoch. This process continued for each of the 30 concentrations, totaling to 60,000 rounds of optimization for the model. Figure 1 shows the results of this training progression for a single concentration. At the end of training we had created a generalized model capable of predicting values for any concentration within a reasonable range. Since these generated samples are being used to improve another model, the metrics for evaluation were mainly comparison between real and generated data to ensure realistic data generation.

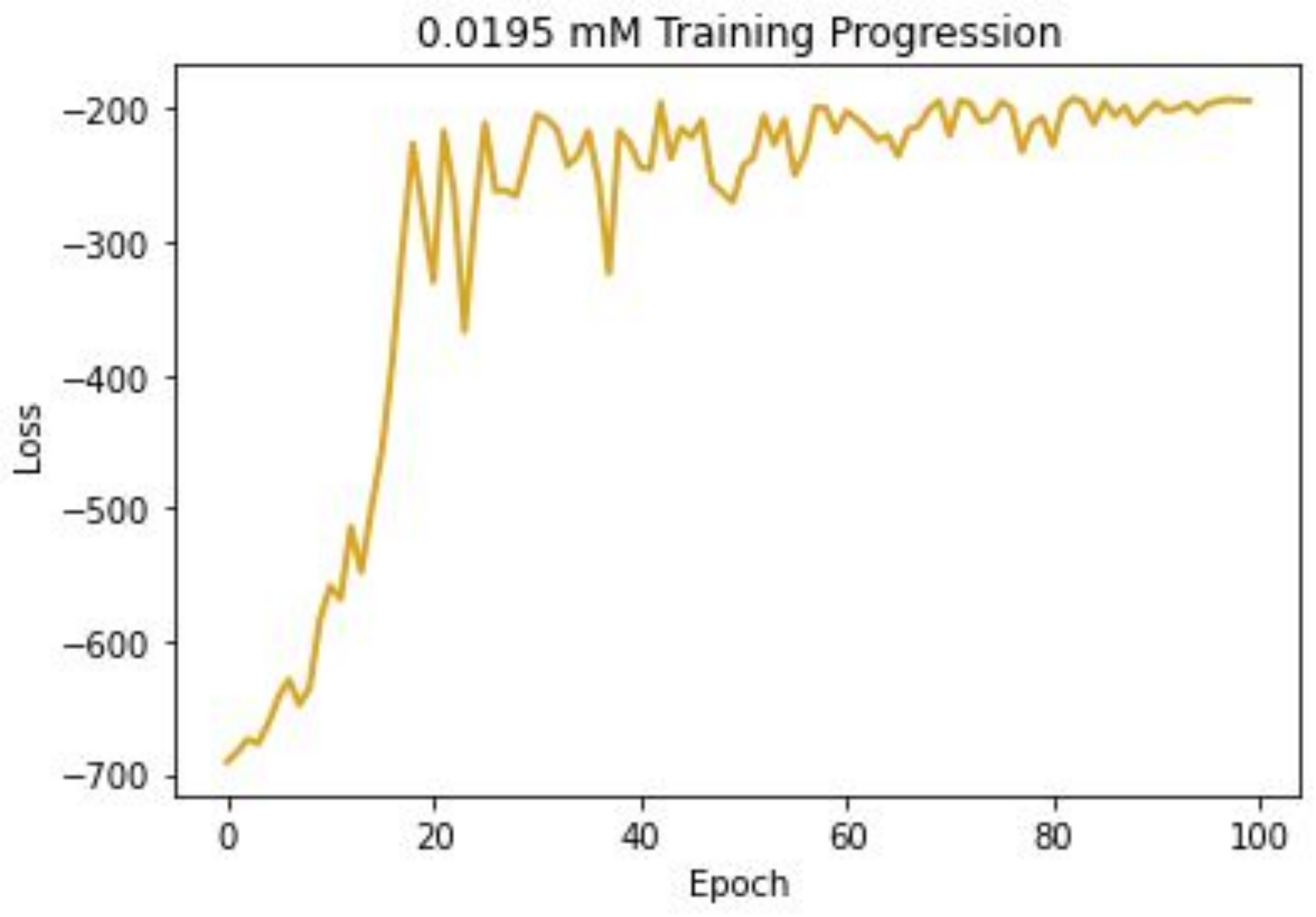

Figure 3 - A line plot of the loss versus the epoch of training. This graph allows us to see the progression in accuracy for a single concentration.

Background

- Phenolic compounds inhibit fermentation process when producing ethanol from biomass.
- Biofuel industry cannot efficiently monitor the concentration of phenolic compounds to optimize the fermentation process.
- Electrochemiluminescence (ECL) sensors are expensive and bulky, inefficient to use in mass.
- We are creating smartphone ECL sensor that can process images or current data to predict the concentration of phenolic compounds.
- Currently real data collection takes time so the training set for the predictive model is sparse, leading to an inflexible model.
- How can we create more realistic current data to augment the training set, making the predictive model more accurate?

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.

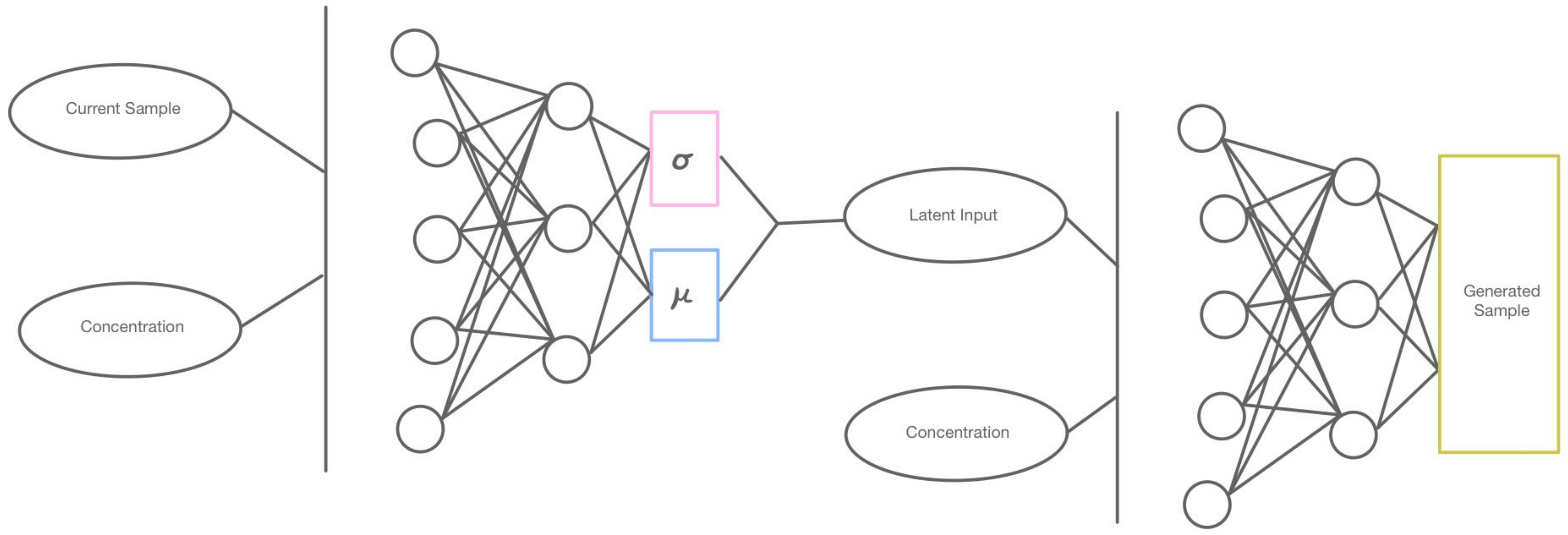

Figure 2 - A diagram of the general variational autoencoder model architecture.

The model trains using a classic training approach, stepping through each epoch applying the Adam optimizer along the way, considering the loss. The most complicated aspect to the model is arguably the loss function. For this we decided to use Evidence Lower Bound in order to measure the loss. We decompose ELBO into the Monte Carlo estimate of the expectation for a single sample, shown in Equation 1. The model continually trains, attempting to minimize the Monte Carlo estimate.

$$\log p(x|z) + \log p(z) - \log q(z|x)$$

Equation 1

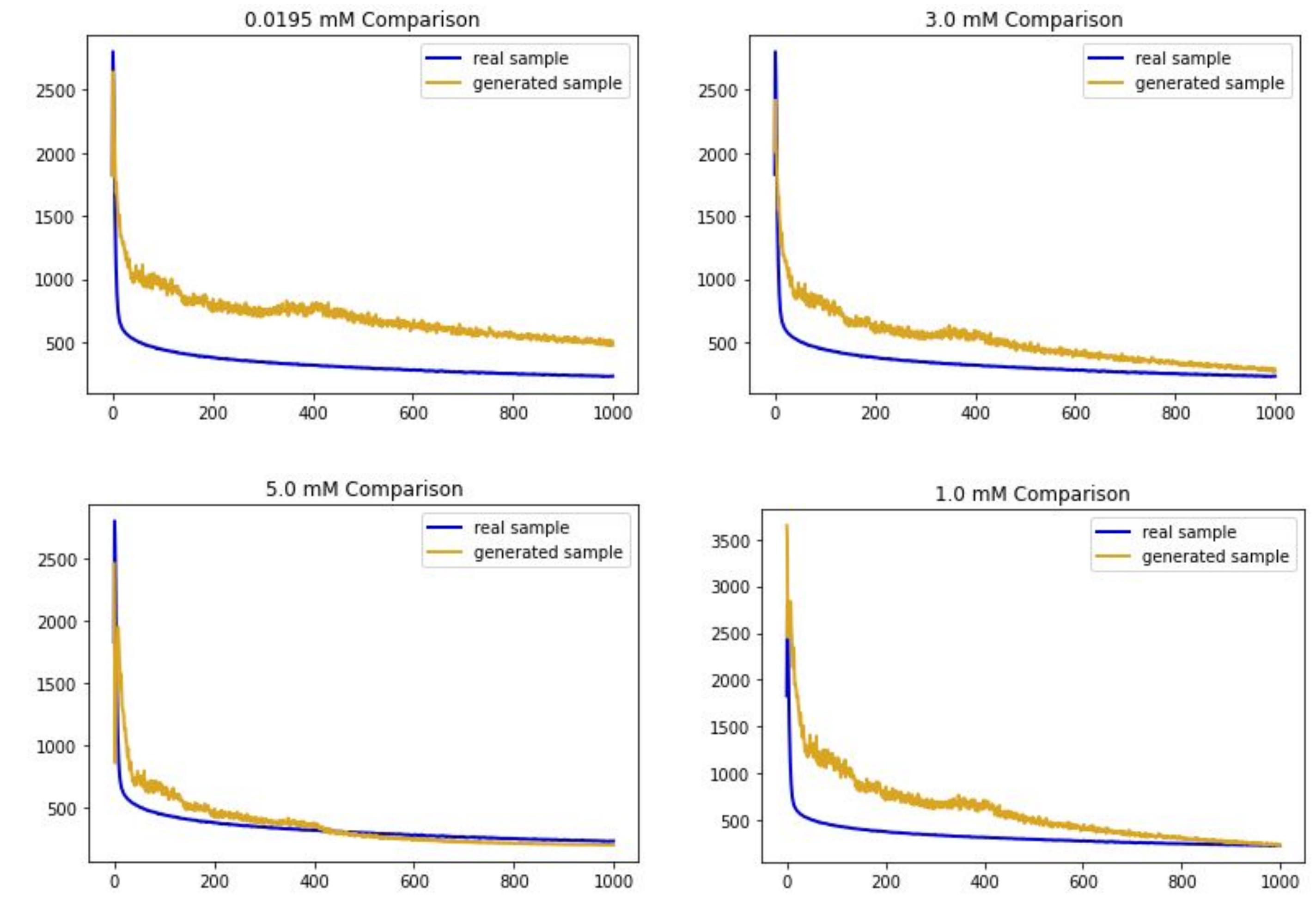

Figure 3 - A comparison between the real curve samples and the generated curve samples.

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.

Selected Bibliography

- "Convolutional Variational Autoencoder : TensorFlow Core." *TensorFlow*, 2021, www.tensorflow.org/tutorials/generative/cvae.
- Diederik P. Kingma and Jimmy Lei Ba. Adam : A method for stochastic optimization. arXiv:1412.6980v9. 2014
- Luo, Yun, et al. "Data Augmentation for Enhancing EEG-Based Emotion Recognition with Deep Generative Models." *Journal of Neural Engineering*, vol. 17, no. 5, 2020, p. 056021., doi:10.1088/1741-2552/abb580.
- Ccopa Rivera, E.; Swerdlow, J.J.; Summerscales, R.L.; Uppala, P.P.T.; Maciel Filho, R.; Neto, M.R.C.; Kwon, H.J. Data-Driven Modeling of Smartphone-Based Electrochemiluminescence Sensor Data Using Artificial Intelligence. *Sensors* **2020**, *20*, 625. https://doi.org/10.3390/s20030625