Abstract

Despite the recent advances in data organization and structuring, electronic medical records (EMRs) can often contain unstructured raw data, temporally constrained measurements, multichannel signal data and image data. Cohort retrieval, the action of finding a group of observations with similar properties, of these signals will allow us to compare and contrast the signals in large quantities We present a proof of concept system that can alleviate this problem by mapping raw data to a compressed 64-dimensional space where the Euclidean distance between data is a measure of similarity. Using electroencephalographs (EEGs) as a case study, we optimize a deep neural network mapping from the spectrogram of EEG data to a latent space by using triplet loss. After this mapping, distance-based methods, such as nearest neighbors search, could be employed to find similar EEG records by treating the embeddings as the keys to the EEG signal in a database as part of a cohort retrieval system. To verify that this method learns a meaningful representation of the data, we apply a six-class k-NN classifier to the output, a binary (seizure-like and noise-like signal) k-NN classifier to the output and visualize the output latent space using the t-SNE dimensionality reduction technique. We achieve a 60.4% six-class signal classification accuracy, a 90.1% binary seizure classification accuracy on the TUH EEG Cohorts dataset and observe distinct clusters in a reduced dimension latent space discovered using the t-SNE algorithm.