Josias Moukpe

Neural Networks

9-26-2022

Project 1 Proposal

It's becoming increasingly evident that synthetic data in Machine Learning is extremely valuable [1]. More and more models will be built and trained with some form of synthetic data to achieve better and more robust performance. However, currently, effective synthetic data is generated by trial and error. There is no system to measure the quality of synthetic data and generate such data with the highest quality. As part of this project, I aim to investigate the ability of Neural Networks to produce high-quality synthetic data and will develop a technique to measure how effective (or in-distribution) a synthetic data point is. By the end of this project, I hope to have developed a method that can measure how in-distribution a synthetic data sample is and an algorithm that can reliably generate in-distribution synthetic data. To achieve this, I plan on leveraging the generator-discriminator adversarial properties of Generative Adversarial Networks (or GANs) to generate new synthetic data points (like a regression task) and discriminate against bad (out-of-distribution) synthetic data points (like a classification task). The exact architecture of the system will be determined during the project.

As part of the data sourcing, I plan on using both labeled and unlabeled datasets. Those datasets will be used for training and validation of the Neural Networks. I will use a variety of datasets, including MNIST, Kaggle datasets, roboflow, etc.

I will be working solo on this project, and I will handle the data, the research, and the development of the project.

References:

[1] https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/?sh=3a7edcaf7523