CSCE 554 Project Proposal

Marvin Newlin

20 Aug 19

The overall goal of this research is to introduce and analyze a set of metrics that can be used to inform the process of generating synthetic log files.

In order to examine the behavior of the metrics, we will take 100 samples, of which a certain number are generated and the rest real and then increase the number of fake samples over time to see what the behavior of the scores is. Ideally the scores will increase as the ratio of real to fake samples increase. Additionally, we plan to run ANOVA on the scores to examine the significance of the scores.

1. What will the response variable be that you are observing?

We will be observing the scores of the different metrics to analyze their behaviors. In particular, we will be looking at the scores of the metrics against the ratios of real versus fake samples scored.

Ideally, the metric’s scores should increase as the ratio of real samples to fake samples increases.

1. What is/are the factor(s) you will be varying?

The factor for this experiment will be the metric scores. Currently, we plan to use the following metrics: Wasserstein distance, Frechet Inception Distance, Cross Entropy Loss, Perplexity, and the Bilingual Language Evaluation Understudy (BLEU) score.

The Wasserstein and Frechet Inception distances are baseline GAN evaluation measures used in image based GANs. We incorporate these distances as a baseline for other distances in order to see whether these “standard” metrics from image based GANs can be used as metrics for text based GANs.

1. Are there any nuisance factors that you will be "blocking"?

No nuisance factors

1. How will you actually collect the data required to run a statistical test?

Using the Unified Host and Network Dataset for this experiment. Generated samples are based on samples from this dataset.