**Advanced Machine Learning - Final Project Proposal**

Yaniv Tal 031431166

Miriam Horovicz 313246704

**Anchor Paper - GP-VAE: Deep Probabilistic Time Series Imputation**

The paper presents a deep probabilistic model for multivariate time series imputation, combining ideas from variational autoencoders and gaussian processes. The VAE maps the missing data from the input space into a latent space where the temporal dynamics are modeled by the GP. They used structured variational inference to approximate the latent GP posterior, which reflects the temporal correlations of the data more accurately than a fully factorized approximation. At the same time, inference in the variational distribution is still efficient, as opposed to inference in the full GP posterior. They empirically validated the proposed model on benchmark data sets and real-world medical data. They observed that the model outperforms classical baselines as well as modern deep learning approaches on these tasks and performs comparably to the state of the art.

**Innovative Part**

We would like to use the model proposed in paper for different task of synthesize high quality time series data. One of the biggest challenges in real world applications is limited data. In many cases, the amount of supplied real data is not enough for training accurate deep learning models. In such cases, methods for synthesizing high quality data are valuable. We will use GP-VAE for synthesize data and show that using it with limited available data improves performance of our downstream task.

Our proposed research flow contains 2 parts: building representative dataset, synthetic data generation & evaluation.

**Build Representative Dataset**

Goal of this part is simulating limited dataset scenario in real life. We will start with full MNIST dataset that we know achieving high accuracy in CNN classification model. Then limiting the data we use, we will get much less accurate results because of lack of data.

We will find such representative limited dataset by following steps:

* Take MNIST data
* Take different sampling sizes (0.5%, 2%, 5%, 10%, 50%, ... , 100%)
* For each different sampling size build CNN classification model and explore accuracy results
* We expect getting graph of decreased accuracy as sampling size decreases
* We will take sampling size that have low accuracy result

By the end of this part, we will have sample MNIST data achieving low results in classification model. This will serve us as limited dataset we would like to improve classification model results on by adding synthetic data as described below.

**Synthetic data generation & evaluation**

Goal of this part is improving our classification model accuracy by adding synthetic data. Synthetic data will create by removing some of the data points from original data and impute them by trained GP-VAE model. This part will be iterative cycles of trying different synthesizing methods, building model and compare results.

Synthetic data generation will be done by following:

* Take each original image
* Remove some fraction of the pixels. We will try different configurations of fraction, spread etc.
* Run GP-VAE to impute and create new samples
* Train classification model on the full data (original + synthetic) and evaluate results
* We assume that model accuracy will improve compared to trained model with original limited data only

By the end of this part, we will show that generating synthetic data with GP-VAE improves accuracy of our downstream classification task.

\*\* If we will find that MNIST doesn’t serve us we may also explore our idea with other datasets i.e., cifar10

**Timeline**

Anchor part:

26.6 – 10.7: Learn the paper with relevant citations

10.7 - 30.7: Reconstruct Papers’ results

Innovative part:

1.7 – 15.7: Build representative dataset

15.7 – 23.8: Synthetic data generation & evaluation

23.8 – 28.8: Final report