# Unifying Transformers and Convolutional Neural Processes

**Name**: Lakee Sivaraya
**College**: Emmanuel College
**Supervisor**: Prof. Richard E. Turner

Neural Processes (NPs) are a class of general purpose models designed for uncertainty-aware meta learning. Simply, NPs learn a probability distribution over the output values, typically characterized by a mean and variance of a Gaussian distribution.

Knowledge of the uncertainty in predictions is crucial for real-world applications such as climate modelling, time series forecasting and decision-making in healthcare. In these applications, the ability to quantify uncertainty is as important as the prediction itself.

Standard DeepSet based NP models exhibit some limitations, particularly its inability to model complicated relationships in high dimensional data. To extend the capabilities of NPs, we replace the DeepSet with two powerful architectures: Convolutional Neural Networks (CNNs) and Transformers giving rise to the Convolutional Neural Processes (ConvNPs) and the Transformer Neural Processes (TNPs) respectively.

Both models are based off two powerful architectures which are bound to have their own strengths and weaknesses. The ConvNP is based off CNNs, which are known for their ability to learn spatial patterns in the data via convolutional filters. The TNP is based off Transformers, which are known for their ability to model long-range dependencies in the data. We compare the ConvNP and TNP on a range of datasets, including smooth, discontinuous and periodic datasets to quantify their performance.

Our experiments demonstrate that the TNP outperforms the ConvNP, especially in higher dimensions. However, TNPs display unpredictable behavior when extrapolating outside the training data, while ConvNPs remain stable. We hypothesize the stability of the ConvNP is due to the *translation equivariance* property of CNNs. Translation equivariance enforces the model to produce the same output when the input is translated, such property is greatly beneficial when modelling stationary or periodic data, e.g. spatio-temporal data. We hypothesize, the lack of translation equivariance in TNPs leads to poor generalization on data far from the training distribution.

To address the limitations of the TNP, we introduce translation equivariance to the TNP, by modifying the self-attention mechanism. We call this model the *Translation Equivariant Transformer Neural Process* (TETNP). Our experiments show the TETNP outperforming the TNP and ConvNP on datasets that exhibit strong discontinuities and periodicity. On smooth datasets, the TETNP performs comparably to the TNP which is expected since these datasets do not benefit much from translation equivariance.

The quadratic memory complexity of full self attention is a key bottleneck in the TNP and TETNP models, limiting their scalability to large datasets. This thesis explores several methods to reduce the memory complexity of the TNP by the use of pseduotokens, and approximations to the self-attention mechanism via kernelization and MLPs. While our experiments demonstrates a substantial reduction in the memory usage, the

approximations come at the cost of performance. All linear models struggled to model the discontinuous datasets, highlighting the loss of expressiveness in the model.

Overall this thesis provides a comprehensive study of the TNP and ConvNP models. We establish the importance of translation equivariance in the TNP and demonstrate its benefits. We showed that the ConvNP scales poorly with dimensionality unlike the TNP and TETNP.

We conclude that the TETNP model is the most promising model, with strong performance across a range of datasets especially in higher dimensions, making them a clear choice for large scale applications.